$\boldsymbol{Q(W) = Q(W|\mu,\rho)}$ 는 매개변수에 기반한 확률분포이므로 경사하강 기법을 이용하여 $\boldsymbol{L(Q)}$의 값을 최대화할 수 있으며, $\boldsymbol{Q(W)}$의 기댓값은 표본 추출을 통한 몬테카를로 근사법을 이용하여 계산한다. 신경망의 가중치를 재매개변수화한 $\mathbf{w} = \mu + log(1 + exp(\rho))^{\circ}\epsilon$ 에서는 정규분포 $\boldsymbol{N(0,1)}$를 따르는 변수 ϵ의 표본을 추출하여 가중치 변수의 표본을 구한다.

2.2. 피셔 정보행렬을 이용한 학습모델

일반적인 신경망을 이용하여 순차적 다중작업 학습을 원활하게 수행하는 방법으로는 [3]에서 제안한 바와 같이 기존의 작업과 상관도가 높은 매개변수의 변화를 상대적으로 억제하는 방법이 있다. 이 때 비용함수에는 로그 우도(log likelihood)에 대한 피셔 정보행렬(Fisher information matrix)의 대각항인 F_i가 포함된 항이 아래와 같은 형태로 추가된다.

$$\boldsymbol{L(W) = L_B(W) + \sum_i \frac{\lambda}{2} F_i (w_i - w^*_{A,i})^2}$$

위 수식에서 $L_B(W)$는 현재 학습중인 작업 B 에 대한 로그 우도를, $w^*_{A,i}$는 작업 A 에 대한 학습이 끝났을 때 해당 매개변수의 값을 나타내며, λ는 작업 A 가 갖는 중요도를 나타낸다. F_i가 크다는 것은 곧 대응되는 w_i의 변화에 따른 로그 우도의 변화가 크다는 것을 의미하므로 해당 매개변수의 중요도가 높음을 의미한다.

2.3. 가중치 분포를 이용한 경사도 조정

베이지안 신경망의 가중치를 재매개변수화한 $\mathbf{w} = \mu + log(1 + exp(\rho))^{\circ}\epsilon$ 에서 μ 는 대응되는 확률분포의 평균값을, $\log(1+\exp(\rho))$ 는 표준편차를 의미한다. 즉 $\log(1+\exp(\rho))$의 값이 작을수록 확률분포에서 추출되는 가중치의 불확실성이 작아지며. 해당 가중치가 과거에 학습한 작업에 대해 갖는 중요성은 반대로 높아진다고 볼 수 있다. 따라서 아래의 수식과 같이 μ에 대한 학습을 진행할 때 적용되는 경사도에 $\log(1+\exp(\rho))$를 추가로 곱하면 학습이 끝난 작업에 대한 성능에 기여도가 높은 매개변수의 변화를 상대적으로 억제할 수 있다. 아래 수식에서 lr은 베이지안 신경망의 학습률을 나타낸다.

$$\boldsymbol{\mu_i \leftarrow \mu_i - lr \times \log(1 + \exp(\rho_i)) \times \frac{\partial L(W)}{\partial \mu_i}}$$

위의 수식을 그대로 베이지안 신경망에 적용하는 경우 초기 학습이 불필요하게 느려지는 현상이 있다. 이를 해결하기 위해 $0 \le \lambda \le 1$ 를 만족하는 λ 를 아래 수식과 같이 도입하여 매개변수의 학습 속도가 지나치게 느려지는 것을 방지할 수 있다. 본 연구에서는 $\lambda = 0.3$ 으로 설정한 상태로 실험을 진행하였다.

$$\boldsymbol{\mu_i \leftarrow \mu_i - lr \times \{\lambda + \log(1 + \exp(\rho_i))\} \times \frac{\partial L(W)}{\partial \mu_i}}$$

3. 실험 결과 분석 및 논의

3.1. 실험 환경

본 연구에서는 MNIST 숫자 필기체 데이터[7]를 실험에 활용하였다. 순차적 다중작업 학습 환경을 구현하기 위해 데이터 내 모든 이미지의 픽셀 순서를 동일하게 뒤섞은 변형본을 만들며, 이러한 변형본 2 종류를 원본 데이터에 순차적으로 추가한 것을 학습 및 성능 검증에 사용하였다. 실험에 사용된 모델은 일반적인 신경망과 베이지안 신경망의 경사도를 조절하여 적용하는 모델의 총 2 종류이다. 각 모델은 순차적으로 MNIST 원본 데이터, 첫 번째 변형본, 두 번째 변형본을 학습하였으며, 각 데이터에 대한 모델의 성능은 처음부터 세 가지 데이터 모두에 대해 측정하여 그래프로 추이를 나타내고 경향을 분석하였다. 모든 신경망은 784 개, 50 개, 10 개 노드가 각 층에 배치된 구조를 사용하였다.

3.2. 실험 결과

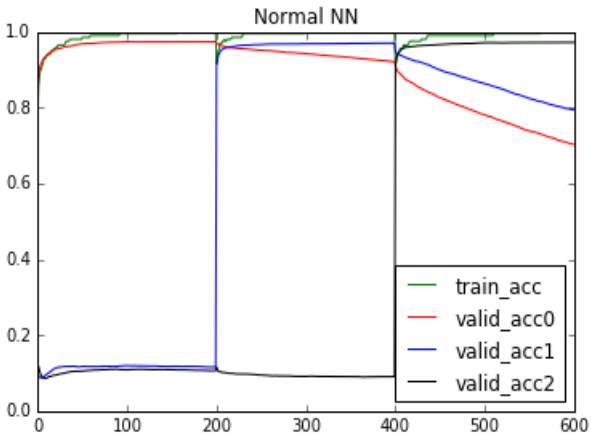

그림 1. 일반적인 신경망을 이용한 실험

그림에서 빨간색 선은 첫 번째로 학습하는 데이터인 MNIST 원본에 대한 성능을, 파란색 선은 두 번째로 학습하는 MNIST 1 차 변형본에 대한 성능을, 검은색 선은 세 번째로 학습하는 MNIST 2 차 변형본에 대한 성능을 나타낸다. 그림 1 과 같이 로그 우도만을 비용함수로 사용한 일반적인 신경망의 경우, 새로운 작업을 학습하는 과정에서 과거에 학습한 작업에 대한 정확도가 지속적으로 낮아지는 현상을 보인다.

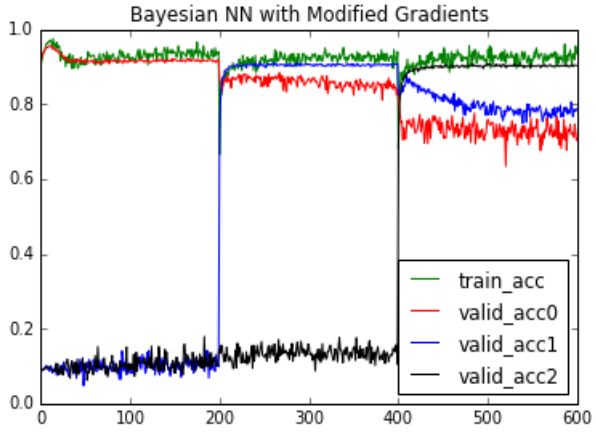

그림 2. 경사도를 조정한 베이지안 신경망 실험

그림 2 는 본 연구에서 제안한 기법을 베이지안 신경망에 적용하여 실험한 결과이다. 이전에 학습한 작업일수록 정확도가 떨어지는 추이를 보이는 것은 일반적인 신경망과 유사하나, 경사도를 조정한 베이지안 신경망에서는 이전 작업에 대한 정확도의 감소세가 일반적인 신경망에 비해 완만하며 일정 수준에서 수렴하는 경향을 보인다.

모델 종류	작업 1 최종성능	작업 2 최종성능	작업 3 최종성능
일반 신경망	70.45%	79.53%	97.17%
베이지안 신경망 +경사도조정	72.54%	78.34%	90.22%

표 1. 실험한 모델의 각 작업에 대한 최종성능

표 1 은 학습이 끝난 상태에서 각 모델이 세 가지 작업에 대해 보이는 정확도를 정리한 것이다. 베이지안 신경망에서 각 작업에 대해 수렴하는 성능은 일반적인 신경망에 비해 낮지만, 이전에 학습한 작업에 대해 보이는 성능의 감소 정도는 일반적인 신경망에 비해 적은 점을 감안하면 베이지안 신경망이 과거에 학습한 작업을 더 잘 기억하는 것으로 볼 수 있다.

4. 결론

본 연구에서는 심층 베이지안 신경망을 이용하여 순차적 다중작업 학습 환경에서 높은 성능을 유지하는 시스템을 제안하였다. 제안된 모델은 MNIST 이미지 원본과 변형본이 순차적으로 주어지는 실험에서 일반적인 신경망을 포함한 기존의 모델에 비해 과거에 학습한 작업을 기억하는 측면에서 더 우수한 성능을 보였다. 한편 베이지안 신경망이 마지막으로 학습한 작업에 대해 보이는 정확도가 일반적인 신경망에 비해 낮아, 베이지안 신경망의 기존 작업 기억능력을 유지하면서도 단일 작업 학습에서의 성능을 개선하는 후속 연구가 필요하다.

5. 감사의 글

이 연구는 국방생체모방 자율로봇 특화연구센터를 통한 방위사업청과 국방과학연구소 연구비(UD130070ID-BMRR) 지원으로 수행되었음.

참고문헌

[1] Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D., “Weight uncertainty in neural network”, In *Proceedings of The 32nd International Conference on Machine Learning*, 1613–1622, 2015.
[2] Fei-Fei, L., Fergus, R. & Perona, P., “One-shot learning of object categories”, *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.
[3] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the National Academy of Sciences* (2017): 201611835.
[4] Opper, M., “A bayesian approach to on-line learning”, In *On-line learning in neural networks*, 363–378, Cambridge University Press, 1999.
[5] Pascanu, Razvan, and Yoshua Bengio. "Revisiting natural gradient for deep networks." *arXiv preprint arXiv: 1301.3584* (2013).
[6] Winther, O. & Solla, S. A. “Optimal Bayesian online learning. Theoretical Aspects of Neural Computation”, (TANC-97), KYM Wong, I. King and D.-Y. Yeung eds. Springer Verlag, Singapore, 1998.
[7] Yann, L., Corinna C., “The MNIST database of handwritten digits”, *URL http://yhann.lecun.com/exdb/mnist/*, 1998.